Gépekkel emberi nyelven

Kontrollált természetes nyelvű interfészek

Mészáros Tamás
http://www.mit.bme.hu/~meszaros/

Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

Problémafelvetés: alkalmazási igények

- „Beszélni akarok a gépekkel, nem gépelni”
  - Más interfészek erősen korlátozottak (mobil, autóvezetés, stb.)
  - Apple Siri, Google Voice, Microsoft Cortana, stb.

- „Nem értem program kezelését. Elmondhatom, mit szeretnék?”
  - Merre van …? Hogy kell eljutni ….? Melyik az a …, amelyik …?
  - Komplex felhasználói felületek használata nehézkes, a betanulás sok idő

- „Nem értem ezt a nyelvet.”
  - SPARQL, PQL, Prolog, XML, RDF, OWL
  - A lekérdezés és a tudásbevitel természetes nyelven a legegyszerűbb.

- „Szeretném, ha a gép lefordítaná nekem ezt a szöveget!”
  - Általam nem értett nyelven közzétett tartalom lefordítása
  - A saját szövegemet más nyelven is szeretném közzéadni.

...

A természetes nyelvű kommunikáció

- A természetes nyelv az emberi kommunikáció meghatározó formája.
  - Mindenki ismeri (persze nem egyfélét, de egyet jellemzően igen).
  - Így osztjuk meg egymással a világról szerzett ismereteinket.
  - Tudásunk leírásának, gondolataink közzétételének alapvető eszköze.
  - A segítségével befolyásolhatjuk a körülöttünk levő világot.
  - Kis kortól kezdve tanuljuk és folyamatosan használjuk is.

- Miért nem kommunikálnak velünk a gépek természetes nyelven?
  (video)
  - Nagyon sok esetben mennyivel egyszerűbb lenne!
    - Bonyolult felhasználói felületek,
    - mesterségesen létrehozott (programozási) nyelvek,
    - gyenge kifejezőerejű vagy nem létező kommunikációs interfészek helyett.
  - Látunk erre példákat, de valahogy nem mindennaposak és nem átütőek.

  Miért?

Természetes nyelvű közlések gépi megértése

- Mi az a természetes nyelv? Mik a szabályai? Hogyan használjuk?
  - A nyelv egy *élő* dolog, meglehetősen összetett rendszer
  - Szabályai időben és térben (beszélőről beszélőre is) változnak
  - A használatához komoly háttértudás szükséges (miről beszélünk?)

- Természetes nyelvű ember-gép interfészek kialakítása
  - Az MI kutatások több évtizedes, klasszikus területe.
  - Sok komoly eredményt tud felmutatni, számos könyvet és publikációt
  - Sokrétű kutatási terület

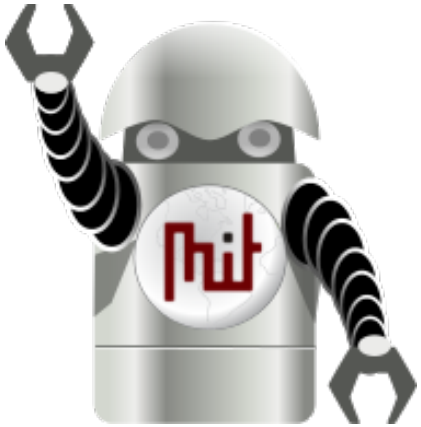
    Beszédfelismerés, szintaktikai elemzés, szemantikai értelmezés, párbeszéd-kezelés, mondatgenerálás, beszédszintézis, gépi fordítás, helyesírás-ellenőrzés, kérdésmegválaszoló rendszerek, stb.
  - Számítógépes nyelvészet (computational linguistics)
  - Eredményeit sokfelé látjuk az iparban és mindennapjainkban
  - Mégis... a gépekkel nem tudunk úgy társalogni, mint az emberekkel.
  - Különösen az értelmezés és a megértés nehéz feladat

Mitől nehéz egy ilyen rendszert készíteni?

- Beszédfelismerés (audio jelek feldolgozása, szöveggé alakítása)
  - Sokféle hang, stílus, beszédhibák - néha még nekünk sem triviális
- A nyelv szintaktikai szabályainak ismerete
  - Mit szabad és mit nem. A mondat részei és szerepeik. Írásjelek (?!).
- Szókincs beépítése
  - Milyen szavakat használhatunk
- Értelmezés
  - Szavak értelme, szavakból épített kifejezések értelmezése
  - A mondatok értelmezése, beleértve az írásjelek módosító hatását is.
- Párbeszédek kezelése
  - Egy közlés értelme alapvetően függhet a korábbiaktól („Nem.”)
- Hivatkozások feloldása
  - „Ő volt az, aki...”
- A válasz előállítása (mit és hogyan mondunk el)
  - Karakterek összefűzése értelmes mondattá
- Az előállított válasz hangjelekké alakítása (kiejtés, hangsúly)

Mégis, mennyire nehéz egy ilyen rendszert készíteni?

- **BME Tibi**: egyszerű kérdésmegválaszoló
  - Android alkalmazás
  - MI fogalmakról és időjárásról lehet kérdezni.



- A program részei
  - Beszédfelismerő: Google
  - Szintaktikai elemző: regexp
  - Értelmező: egyszerű szabályalapú mintaillesztő
  - Válasz előállító: AsyncHTTPClient + MI Almanach, Köpönyeg, Időkép
  - Beszédszintetizátor: Android TTS + Mariska hang

- A fejlesztés menete
  - 1. nap: Android beszédfelismerő alkalmazás
  - 2. nap: regexp elemző, szabályalapú értelmező és beszédszintetizátor
  - 3. nap: háttérkiszolgálók beépítése

Kutatási területem: kontrollált természetes nyelvek

- Alapötlet: a feldolgozás és megértés nehézségeinek eliminálása
  - Rögzítsük a nyelvtani szabályokat és a szókincset
  - Tegyük egyértelművé a nyelv értelmezését
  - Máris megoldottuk a legnagyobb problémákat

- Sajnos generáltunk újabb problémákat
  - Hogyan hozzuk létre őket?
  - Milyen nyelvtani konstrukciókat engedünk meg?
  - Milyen nyelvtani elemzőt használjunk?
  - Mekkora legyen a szókincs?

  - Hogyan vegyük rá a felhasználót a szabályok betartására?
  - Honnan tudja a felhasználó, mi része a szókincsnek?

Kontrollált természetes nyelv

- Definíció

> *A controlled language (CL) is a restricted version of a natural language which has been engineered to meet a special purpose, most often that of writing technical documentation for non-native speakers of the document language. A typical CL uses a well-defined subset of a language's grammar and lexicon, but adds the terminology needed in a technical domain. (Kittredge, 2003)*
>
> *Controlled natural language is a subset of natural language that can be accurately and efficiently processed by a computer, but is expressive enough to allow natural usage by non-specialists. (Fuchs and Schwitter, 1995)*

A kontrollált nyelv (controlled language, CL) egy olyan mesterséges nyelv, amely a sikeres számítógépes feldolgozhatóság érdekében szűkíti egy természetes nyelv nyelvtani szabályait, szókincsét és szemantikáját megőrizve annak természetes jellegét.

- Honlapok

https://sites.google.com/site/controllednaturallanguage/

http://attempto.ifi.uzh.ch/site/cnl2012/

http://project.mit.bme.hu/clif/biblio (alkalmi gyűjtés 2009-ből)

Példák kontrollált nyelvekre

- Basic English (1932), EasyEnglish (1997), Wikipedia Simple English
  - A kifejezőkészség javítása érdekében a globális (többnyelvű) világban
- INTELLECT natural language database query system (1981), ...
  - Adatbázisok egyszerűbb lekérdezése természetes nyelven
- Caterpillar, Airbus, Boeing, ... technical languages (1973-), 40+ féle
  - felhasználói, karbantartási kézikönyvek, később (1996-) gépi fordítás
- FAA Air Traffic Control (2010), AECMA/ASD technical english (1980)
  - Az előbbi nyelvek „hatósági” szabványosítása
- “Massachusetts Legislative Drafting Language” (2003)
  - a törvénytervezetek egységes, könnyebben értelmezhető leírására
- Attempto Controlled English (ACE, 1995, 2008), CLCE (Sowa, 2004)
  - szoftverspecifikáció, tudásleírás, logikai tudásreprezentáció megfelelője
- CLOnE (2007), GINO (2006), Ginseng (2006)
  - korlátozott ontológiaszerkesztésre és lekérdezésre
- ... ezernyi más, l. „Kuhn: Survey and Classification of CNLs” + ábra

Attempto Controlled English (ACE)

- Az angol nyelv egy részhalmaza
  - egyes- és többesszám kezelése (a man, some men)
  - egzisztenciális és univerzális kvantorok (there is a man, every man)
  - számosság (5 men, at least 5 men, less than 5 men)
  - névmások, vonatkozó és általános is (a man that has a cat, everybody)
  - negálás, konjunkció, diszjunkció, utaló hivatkozások főnevekere, stb.
- A mondatok szerkezete
  - Deklaratív mondatok (Every man has a dog.)
  - Változókkal kiegészítve (There is a man X who has a dog.)
  - Eldöntentő kérdések (Does Peter have a dog?)
  - Wh-kérdések (Who has a dog? Peter has what?)
- Elsőrendű logikai állításokra képezhetők le a mondatai
  - Az eszközkészletében található elemző (APE) végzi el a leképezést.
  - A következtető gépe képes így kialakított tudásbázisok használatára
- Kipróbálható: http://attempto.ifi.uzh.ch/ape/
- Integrálható: http://technologies.kmi.open.ac.uk/aqualog/

Ontológia szerkesztés: CLOnE, GINO

- A szemantikus publikálás megköveteli ontológiák kidolgozását
  - Az RDF/OWL egyre szélesebb alkalmazói körrel bír, de
  - a szerkesztőeszközök (pl. Protege) használata nem triviális.

- A kontrollált nyelvek (Rabbit, CLCE, ACE-OWL, PENG-OWL, SOS,...)
  - lehetővé tehetik laikus felhasználók számára is az ontológiaszerkesztést
  - használatuk jóval kevesebb betanulást igényel, mint egy OWL szerkesztő

- CLOnE: Controlled Language for Ontology Editing (2007)
  - A GATE keretrendszert használja a mondatok elemzésére
  - Nyelvi elemzője laza: kulcs kifejezéseket próbál megtalálni és értelmezni

- GINO: Guided Input Natural Language Ontology Editor (2006)
  - alkalmas lekérdezésre (SPARQL nyelvre fordít), és szerkesztésre is
  - inkrementális nyelvi elemző segíti a kontrollált nyelv használatát
  - a Jena Semanti Web keretrendszert használja

Kontrollált nyelvű lekérdezés

- A természetes és a formális nyelvű lekérdezők közé épít hidat
  - A felhasználók által könnyebben elsajátítható, és
  - a formális lekérdező nyelvekre egyértelműen lefordítható
- Adatbázisok lekérdezése („a klasszikus”)
  - Általános természetes nyelvű lekérdezőként indult a 60-as években
  - Kommerciális termékek is megjelentek:
    INTELLECT, IBM LANGUAGEACCESS, stb. (10+)
  - Nem kontrollált, vagy nem pontosan ismert nyelvek
- Ontológiák lekérdezése
  - A szerkesztők „melléktermék”, folytatása, előzménye
  - QuestIO (2008): a CLOnE lekérdező rendszere, SPARQL-re fordít
  - A GINO rendszer lekérdezésre is alkalmas
- Tudástárak lekérdezése jellemzően *alany-reláció-tárgy* kapcsolatokra
  - Általános célú, logikai-alapú nyelvek környékén (ACE, CLCE, stb.)
    - Pl.: AquaLog (2004), FREyA (2010)
  - GAPP (2003): „Foundational Model of Anatomy” tudásbázis lekérdezése
  - ...

Esettanulmány: főnévi vonzattár

(videó)

Esettanulmány: a főnévi vonzattár nyelvtana

'QS' => 'QW NP1 REL NP2',

'QW' => 'melyik',

'NP1' => 'főnév',

'REL' => 'rendelkezik | nem rendelkezik',

'NP2' => 'vonzattal | VJ vonzattal',

'VJ' => 'MJ | VT | MJ VT',

'VT' => 'birtokos | ban,ben | ról,ről | hoz,hez,höz | ból,ből | vhonnan | vhova | vmikor | val,vel | tól,től | ért,miatt | nak,nek | ra,re | vmilyen | vmennyi | n | után | között | ellen | szemben | mellett | belül | felett | iránt | be | felé',

'MJ' => 'MJ1 | MJ2 | MJ1 MJ2',

'MJ1' => 'fakultatív | kötelező',

'MJ2' => 'élőre vonatkozó | élettelenre vonatkozó'

Gépi fordítás támogatása

- A kontrollált nyelvek alkalmazása javíthatja
  - a forrásnyelv érthetőségét, megfogalmazását, és ezáltal
  - a kimeneti nyelvre fordított szöveg pontosságát és érthetőségét is.
- Az alkalmazás módszerei
  - Lexikai kontroll: csak egy előre meghatározott szókincset fogad el
  - Nyelvtani kontroll: csak meghatározott nyelvtani szerkezeteket fordít
  - Összefoglalva: alapvetően nyelvtani és lexikai többértelműség kiszűrése
- Alkalmazások
  - Technikai dokumentáció automatikus gépi fordítása („klasszikus”)
  - Termékinformációk fordítása
- Példák
  - KANT Controlled English (1995)
  - DRAFTER (1996): többnyelvű használati útmutatók készítése
  - CLOUT nyelv, Uwe Muegge munkái
  - MOLTO: Multilingual Online Translation (2010)
- Lásd még: http://www.mt-archive.info/srch/subjects.htm

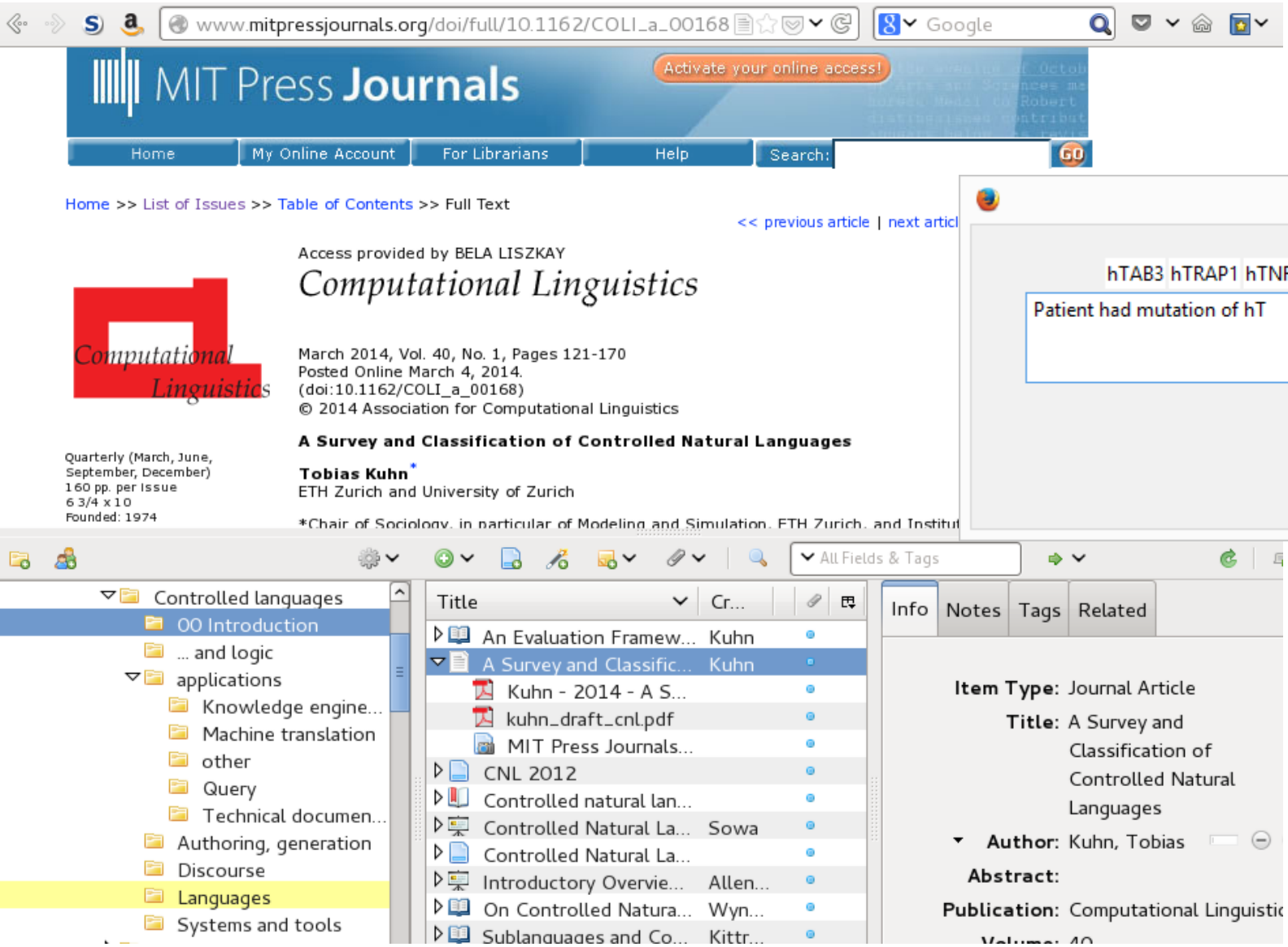
Szemantikus publikálás: CLANN

- A szemantikusan jelölt tartalom létrehozása is igényel támogatást
  - A kidolgozott ontológiák és tartalomnyelvek (XML) használatát segítendő
  - A cél egyrészt a meglevő szövegek szemantikus annotációja,
  - másrészt állítások megfogalmazása a tartalomnyelv szabályai szerint.

- CLANN: Controlled Language for Annotation (2010)
  - Tárgyterületi ontológia + RDF tartalomnyelv + grafikus szövegszerkesztő
  - A CLOnE nyelvre épül, a cél jegyzőkönyvek és riportok annotálása
  - A GATE keretrendszert használja nyelvi elemzési feladatokra
  - Demo: Semantic MediaWiki kiterjesztés

- (Esettanulmány: BME MIT OTKA projekt terv)

Esettanulmány: CNL kivonatok a Zoteroban



A kontrollált nyelvű bevitel támogatása

- Problémák a kontrollált nyelv alkalmazása során
  - Hogyan vegyük rá a felhasználót a szabályok betartására?
  - Honnan tudja a felhasználó, mi része a szókincsnek?

- Lehetséges megoldások
  - Megtanítjuk a felhasználót a szabályokra és a szókincsre
  - A bevitel során nincs támogatás, de az elemzés korrigál, visszajelez
  - Valamilyen beviteli támogatást alkalmazunk

- A szövegbevitel támogatása
  - Az egyszerű menüalapú kiválasztástól
  - a mondatban szereplő fogalmak, konstrukciók grafikus kiválasztásán át
  - egy szövegrészlet lehetséges bővítéseit dinamikusan generáló rendszerig
  - Ez utóbbi új követelményeket támaszt a nyelvi elemzővel szemben.

Összefoglalás

- Problémafelvetés
  - A természetes nyelvű kommunikáció kívánatos, de nehéz a gépeinknek
- A kontrollált nyelv
  - Igyekszik az NLP nehézségeit megszüntetni
  - Rögzíti a nyelvet (amely lehet általános vagy alkalmazásspecifikus)
  - Tanulni kell a használatát vagy valamilyen segítség kell a felhasználónak
- Kontrollált nyelvek
  - Gépi vagy emberi felhasználásra
  - Általános (ACE) vagy speciális céllal (...)
- Alkalmazási területek
  - Adatbázisok és tudásbázisok (ontológiák) lekérdezése
  - Kontrollált tartalomszerkesztés (dokumentálás, fordítás, stb.)
  - Tudásbevitel (logikai kötődésű nyelvek)

Mészáros Tamás, BME MIT

Elérhetőségek

Iroda: 1117 Budapest, Magyar tudósok krt. 2. I. ép. IE437
Tel.: +36 1 463-2899
Fax: +36 1 463-4112
Email: meszaros (*) mit * bme * hu
Személyes honlap: http://home.mit.bme.hu/~meszaros/

Bemutatkozás

Oktatóként, kutatóként, és nem utolsó sorban alkotó mérnökként dolgozom a tanszéken.
Oktatási munkám az informatikus képzésekre koncentrál, az operációs rendszerek és az intelligens rendszerek szakirányok oktatásában veszek részt.
Kutatási területeim: tudásalapú rendszerek, szövegelemzés és -bányászat, kontrollált természetes nyelvű felhasználói felületek, valamint mindezek alkalmazásai az információszolgáltatásban és beszerzésben.
Mérnökként a fentieken kívül webes és XML technológiákkal, valamint Linux rendszerekkel foglalkozom.

Aktuális: Digitális bölcsészet konferenciát szervezek ősszel az MTA-n.

Europass CV
Önálló labor témáim összefoglalása (2015)

Aktuális témakiírásai:

- Villanyspenót bővítések (szövegbányászat, web)
- Történeti szövegek számítógépes elemzése
- Tudásbevitel kontrollált természetes nyelven
- Természetes nyelvű felület Androidon
- Közlekedés modellezése többágens-rendszerrel
- Természetes nyelvű felhasználói felületek

http://www.mit.bme.hu/~meszaros/